dr hab. Paweł Grygiel
Instytut Pedagogiki
Uniwersytet Jagielloński
ul. Batorego 12, 31-135 Kraków
e-mail: pawel.grygiel@uj.edu.pl

Rzeszów, 24.11.2019

**Recenzja rozprawy doktorskiej mgr Radosława Wujcika**

***Zastosowanie teorii odpowiadania na pozycje testowe (IRT) i modeli diagnostycznych w celu poprawy trafności testów statystycznych w diagnozie psychiatrycznej dzieci i młodzieży. Na przykładzie narzędzi do diagnozy ADHD i ASD***

**Promotor: dr hab. n. med. Tadeusz Pietras, prof. UM w Łodzi**
**Promotor pomocniczy: dr Łukasz Mokros**

Przedłożona do oceny dysertacja – wraz z załącznikiem zawierającym polecenia zastosowane w pakietach środowiska R – liczy 230 stron. Praca składa się z trzech rozdziałów oraz podsumowania. Część pierwsza (zatytułowana: „Wstęp teoretyczny") obejmuje: (1) opis podstawowych zagadnień związanych z diagnozą psychiatryczną, z akcentem położonym na rolę jaką w tym procesie odgrywają testy psychologiczne; (2) omówienie podstaw matematycznych oraz różnic pomiędzy klasyczną teorią testów (KTT), teorią odpowiadania na pozycje testowe (IRT) oraz modelami diagnostycznymi (CDM); (3) przybliżenie standardów diagnozy nadpobudliwości psychoruchowej z deficytem uwagi (ADHD) oraz diagnozy zaburzenia ze spektrum autyzmu (ASD). W rozdziale drugim przedstawiono cele pracy oraz hipotezy, grupę badaną, opis narzędzi badawczych oraz procedurę badania. Rozdział trzeci obejmuje z kolei wyniki przeprowadzonych analiz. Przedstawione są one osobno dla dwóch narzędzi będących przedmiotem zainteresowania Autora, to jest: (1) Kwestionariusza do Diagnozy Autyzmu (ASRS) oraz (2) Kwestionariusza do Diagnozy ADHD (Conners 3). W obu przypadkach obejmują one zagadnienia: (1) estymacji i trafności modeli IRT; (2) analizę zróżnicowanego funkcjonowania pozycji testowej (DIF); (3) estymację i trafność modeli IRT; (4) estymację i trafność modeli diagnostycznych (CDM). Rozprawę kończy podsumowanie składające się z dyskusji wyników oraz wniosków.

W części teoretycznej, bo za taką uznaję „Wstęp" Autor bardzo wszechstronnie zarysował tło teoretyczne badanych przez siebie zagadnień, poczynając od przedstawienia typów narzędzi diagnostycznych, zarysowania problemów metodologicznych towarzyszących pomiarowi cech psychicznych, właściwości psychometryczne testów (rzetelność, trafność, stronniczość testów, normy wyniku testowego), poprzez szczegółowe przedstawienie modeli wyniku testowego (klasycznej teorii testów – w tym ujęcie rzetelności w perspektywie KTT, właściwości pozycji testowej, w tym trudności i mocy dyskryminacyjnej]; teorii odpowiadania na pozycje testowe – w tym podstaw teoretycznych IRT, różnych modeli IRT, wielogrupowych modeli IRT, zagadnienie szacowania modeli tego typu, specyfiki rozumienia rzetelności w ramach IRT, DIF w przestrzeni modelowania IRT]; poznawczych modeli diagnostycznych – w tym założeń ogólnego modelu diagnostycznego – GDM oraz modelu diagnostycznego – GDINA), aż po problematykę diagnozy zaburzeń ze spektrum autyzmu (w tym kryteria diagnostyczne ASD, metody używane w diagnozie ASD, opis Zestawu Kwestionariuszy do Diagnozy Autyzmu – ASRS) oraz diagnozy nadpobudliwości ruchowej z deficytem uwagi (w tym kryteria diagnozy ADHD, metody używane w diagnozie ADHD, przedstawienie Zestawu Kwestionariuszy do Diagnozy ADHD – Conners 3).

Do tej części rozprawy można mieć tylko niewielkie zastrzeżenia. Pierwsze wiąże się z opisem ograniczeń związanych ze współczynnikiem zgodności wewnętrznej alfa Cronbacha. Jak wiadomo, klasyczna alfa liczona jest na bazie macierzy korelacji r Pearsona – niezbyt nadaje się więc do szacowania rzetelności testów bazujących na itemach porządkowych. W przypadku najczęściej stosowanych skal badawczych optymalnym rozwiązaniem jest oparcie współczynnika na macierzy korelacji polichorycznej lub (w przypadku zmiennych dychotomicznych) tetrachorycznej. Ponadto zakłada ona (o czym wspomina zresztą Autor), że opiera się ona na założeniu tau-ekwiwalentności – co jest trudne do osiągnięcia w praktyce. Spełnienie tego założenia można zresztą testować w modelowaniu czynnikowym czy analizach IRT (oba sposoby analizy są zresztą wzajemnie „przekładalne"). Ponadto, istnieją współczynniki (jednym z nich jest omawiany przez Autora współczynnik omega), które nie zakładają, że wszystkie pozycje powinny mierzyć cechę w takim samym stopniu. Brakuje mi więc więcej informacji o rzetelności w powiązaniu z poziomem pomiaru oraz różnymi metodami estymacji wyników.

Po drugie, w nazbyt wąski sposób – moim zdaniem – potraktowano także analizę czynnikową, jedną z najstarszych i najlepiej rozwiniętych metod statystycznych służących do badania relacji między wskaźnikami a cechami ukrytymi. Pojawia się ona tylko w kilku miejscach (metoda testowania trafności, analiza homogeniczności) w bardzo ograniczonym

treściowo zakresie. Wydaje się, że przedstawienie specyfiki tego rodzaju modelowania znacząco poszerzyłoby horyzont metodologiczny rozprawy. Analiza czynnikowa pozwala wszak na szacowanie współczynników rzetelności wywodzących się tak z KTT (np. współczynnik omega), jak występujących w ramach IRT (krzywe charakterystyczne opozycji – IRT). Umożliwia wykorzystanie wielu różnych metod szacowania zróżnicowanego funkcjonowania pozycji testowej (tak międzygrupowej, jak wzdłużnej). Analizę struktur czynnikowych można wykorzystać tak do testów wykorzystujących itemy mierzone na skali porządkowej (choć także nominalnej), jak ilościowej). Umożliwiają one również analizę modeli wielowymiarowych, w tym struktur bardziej złożonych, np. podwójnego czynnika czy wyższego rzędu. W konsekwencji wartościowe byłoby także porównanie modelowania czynnikowego z modelowaniem IRT. Generalnie są one wszak „przekładalne", można tak przekształcić parametry modelu czynnikowego, aby otrzymać parametry modelu IRT – i odwrotnie. Występujące między nimi różnice wynikają z odmiennych sposobów parametryzacji cechy ukrytej. Rozwiązania czynnikowe posługują się w tym celu najczęściej (choć nie zawsze) analizą korelacji–kowariancji, właściwą dla metod niepełno-informacyjnych, IRT zaś – metodami pełno informacyjnymi.

Zdaję sobie oczywiście sprawę, że celem pracy nie jest wyczerpujący opis analizy właściwości psychometrycznych, że służyć mają one zarysowaniu metodologicznego tła niezbędnego do zrozumienia analiz właściwych, a przedstawiony materiał jest i tak obszerny i treściwy. Wymiecione uwagi uważam więc za „błahe".

W części dotyczącej metody Doktorant sformułował główny cel badania, jakim jest wypracowanie rozwiązań pozwalających poprawić trafność testów psychologicznych stosowanych w diagnozie psychiatrycznej dzieci młodzieży. Główny cel badania zamierza osiągnąć poprzez: (1) porównanie użyteczności modeli IRT wg trzech metod estymacji (GRM, GPCM i GRSM); (2) porównanie użyteczności modeli IRT wg liczby parametrów (1PL oraz 2PL); (3) analizę DIF ze względu na płeć; (4) analizę DIF ze względu na wiek; (5) ocenę korzyści płynących z wykorzystania modelu wielogrupowego dla płci (MG-IRT); (6) ocenę korzyści płynących z wykorzystania modelu wielogrupowego dla wieku (MG-IRT); (7) porównanie użyteczności dwóch typów modeli diagnostycznych (GDM i GDINA); (8) ocenę CDM w tworzeniu wyniku ogólnego (wyższego rzędu). Związane z tymi celami hipotezy głosiły, że: (1) zastosowanie modeli IRT przyczyni się do poprawy trafności obu narzędzi badawczych; (2) zastosowanie modeli diagnostycznych przyczyni się do poprawy trafności obu narzędzi badawczych; (3) wystąpi zjawisko DIF ze względu na płeć i wiek w przypadku obu analizowanych kwestionariuszy; (4) wykorzystanie wielogrupowych modeli IRT przyczyni się

do poprawy trafności analizowanych skal. Cele pracy sformułowane został jasno i dotyczącą zagadnień ważnych tak z praktycznego (diagnoza psychiatryczna), jak teoretycznego (analiza właściwości psychometrycznych) punktu widzenia. Być może sformułowane w tym miejscu hipotezy należałoby syntetycznie uzasadnić. Uzasadnienie znajduje się wprawdzie we wcześniejszych partiach tekstu (lub z nich wynika), jednak dodanie kilku zdań w tym właśnie miejscu zwiększyłoby przejrzystość argumentacji Autora. Wrażenie (pozytywne) robi wielkość próby badawczej (kalibracyjnej – ponad 1000 osób) oraz fakt wykorzystania do analizy trafności kryterialnej stosunkowo licznej grupy osób z diagnozą ASD oraz ADHD (między 50 a 200 w zależności od narzędzia).

W analizach - przeprowadzonych w zdecydowanej większości w programach działających w otwartym środowisku R - zastosowano adekwatne do celów badania metody statystyczne. Pod względem analitycznym pracy niczego zarzucić nie można.

Drobne uwagi dotyczą wskaźników dopasowania modeli. Wykorzystane zostały przede wszystkim miary bezwzględnego dopasowania. Brakuje miar relatywnego dopasowania, takich choćby jak TLI czy CFI. Generalnie, analizy poszerzyłbym o weryfikację jednowymiarowości wykorzystywanych wymiarów w oparciu o konfirmacyjną analizę czynnikową (w oparciu o macierz korelacji polichorycznych i estymator WLSMV, dostępne (o czym Doktorant z pewnością wie) w pakiecie lavaan lub – komercyjnym – Mplus). Biorąc jednak pod uwagę uzyskane wartości RMSEA nie sądzę, żeby ich wykorzystanie mogło wpłynąć na wnioski. Wykorzystanie modelowania czynnikowego dałoby dodatkowy „zysk", umożliwiając określenie rzetelności nie tylko w oparciu o wielkość błędu standardowego, lecz także np. o współczynnik omega. Do analizy trafności zewnętrznej wykorzystać można byłoby szerszą gamę narzędzi badawczych – choćby w wersji skróconej.

Mam także jedną generalną uwagę związaną z przeprowadzonymi analizami. W pracy brakuje przekonującego uzasadnienia, dlaczego wykorzystano akurat modele jedno-wymiarowe, skoro istnieją wielowymiarowe modele IRT – a badane narzędzia są wielowymiarowe (w przypadku ASRS są to trzy wymiary; zaś Conners-3 siedem wymiarów (przy czym, analizie poddano dwa). Wykorzystanie modeli wielowymiarowych, w tym modelowania struktur wyższego rzędu (np. podwójnego czynnika), umożliwiłoby estymację natężenia cechy generalnej („ogólnego" poziomu ADHD czy ASD). W efekcie można byłoby sprawdzić, czy oszacowanie czynnika głównego nie poprawia trafności pomiaru.

Tak, czy inaczej wnioski z przeprowadzonych analiz są bardzo interesujące. Najlepiej dopasowane są modele GRM, stąd ważna (poprzedzona pogłębiona analizą alternatywnych wyjaśnień odnoszących się do matematycznych właściwości różnych modeli) konkluzja, że

„modele dwuparamatryczne dają większą szansę na poprawienie właściwości psychometrycznych testu)" (strona 194). Analizy wskazują także, że modele GRM dostarczają najbardziej (spośród modeli testowanych) rzetelnych wyników. Znaczące są także wyniki związane z trafnością, choć nie dają jednoznacznych rozwiązań. W przypadku kwestionariusza ASRS zastosowanie modeli IRT nie przyczyniło się do wzrostu poprawności klasyfikacji do porównywanych grup w porównaniu ze wskaźnikami liczonymi zgodnie z podejściem KTT (zwiększyła się czułość, lecz spadła specyficzność). Z kolei w przypadku Conners-3 zastosowanie modelowania IRT spowodowało wzrost czułości testu, przy zachowaniu zbliżonego poziomu specyficzności. Interesujące, że wykorzystywanie IRT nie zmienia układu relacji dla poszczególnych skal w porównaniu z modelowaniem KTT. W efekcie Doktorant stawia następująca (i – w świetle przedstawionych analiz – trafną) diagnozę: „zastosowanie modeli IRT pozwala na poprawę trafności testu stosowanego w diagnozie zaburzeń, lecz korzyści, jakie z niej płyną nie są bardzo duże" (strona 196-197). Zwraca także (nader słusznie) uwagę, że „zyski" płynące z modelowania IRT mogą wiązać się z określonymi charakterystykami konkretnych narzędzi. Pierwsza z postawionych hipotez znalazła więc jedynie częściowe potwierdzenie.

Przedstawione w tym miejscu pracy analizy skłaniają mnie jednocześnie do postawienia dwóch (pośrednio tylko związanych z prowadzonymi przez Doktoranta analizami) pytań: (1) czy i w jaki sposób zastosowanie modeli IRT (estymowanie wyników testu modelowaniem IRT) może przełożyć się na tworzenie „lepszych" norm dla narzędzi badawczych; (2) czy modele IRT mogą zostać zastosowane w praktyce diagnostycznej? W tym drugim przypadku nie chodzi o ocenę właściwości psychometrycznych (na to pytanie Autor sumiennie odpowiedział), lecz ocenę natężenia danej cechy konkretnego dziecka przez diagnostę przeważnie posługującego się sumowaniem/uśrednianiem wartości poszczególnych pozycji testowych dla uzyskania ostatecznego wyniku, co nie ma wiele wspólnego z modelowaniem IRT?

Przeprowadzone analizy wskakują jednocześnie, że oba badane narzędzia obarczone są zróżnicowanym funkcjonowaniem pozycji testowych. Przy czym, siła efektu DIF była mała, oznaczając, że nie wpływa znacząco na estymację wyników. Istotne, że Autor wyjaśnia potencjalną paczynę DIF w przypadku konkretnych pozycji testowych. Potwierdziła się więc trzecia z postawionych hipotez.

Wyniki dowodzą jednocześnie, że zastosowanie wielogrupowych modeli IRT umożliwiających niwelowaniu wpływu DIF na psychometryczne właściwości badanych kwestionariuszy nie poprawia trafności kryterialnej (poprawności klasyfikacji do grupy

klinicznej), choć rzetelność tego typu modelowania była wyższa w przypadku chłopców i dzieci młodszych. Hipoteza czwarta znalazła więc jedynie częściowe potwierdzenie.

Przedstawione dane nie pozwalają jednocześnie na przyjęcie hipotezy drugiej, a więc związanej z modelami diagnostycznymi. Ich wykorzystanie nie poprawia trafności badanych narzędzi diagnostycznych. Jest to wynik interesujący i ważny, biorąc pod uwagę, że modele te są nader rzadko wykorzystywane w Polsce (nawet w badaniach edukacyjnych).

Podsumowując, według mojej oceny, rozprawa doktorska mgr Radosława Wujcika stanowi samodzielne i oryginalne osiągnięcie naukowe. Doktorant wykazał się dużą wiedzą w zakresie analiz psychometrycznych jako takich, ze szczególnym, wszakże uwzględnieniem narzędzi służących do diagnozy psychiatrycznej. Zaplanował i przeprowadził zaawansowane analizy statystyczne na reprezentatywnej próbie dzieci i młodzieży, przez co uzyskane wyniki posłużyły do sformułowania znaczących wniosków. Żadne z podniesionych przez mnie uwag nie ma charakteru zasadniczej natury.

Uważam, że praca „Zastosowanie teorii odpowiadania na pozycje testowe (IRT) i modeli diagnostycznych w celu poprawy trafności testów statystycznych w diagnozie psychiatrycznej dzieci i młodzieży. Na przykładzie narzędzi do diagnozy ADHD i ASD" spełnia kryteria stawiane rozprawom doktorskim w dziedzinie nauk medycznych, stąd wnoszę do Rady Wydziału Lekarskiego Uniwersytetu Medycznego w Łodzi o dopuszczenie jej autora do dalszych etapów przewodu doktorskiego. Jednocześnie, biorąc pod uwagę nowatorski w polskiej literaturze przedmiotu charakter pracy oraz jej wysoką wartość poznawczą i praktyczna, mam przyjemność wnioskować o uznanie jej za wyróżniającą.

dr hab. Paweł Grygiel
Uniwersytet Jagielloński